

Error Free Perfect Secrecy Systems

Siu-Wai Ho, Terence H. Chan, Alex Grant and Chinthani Uduwerelle

Institute for Telecommunications Research

University of South Australia

Abstract

Shannon's fundamental bound for perfect secrecy says that the entropy of the secret message cannot be larger than the entropy of the secret key initially shared by the sender and the legitimate receiver. Massey gave an information theoretic proof of this result, however this proof does not require independence of the key and ciphertext. By further assuming independence, we obtain a tighter lower bound, namely that the key entropy is not less than the logarithm of the message sample size in any cipher achieving perfect secrecy, even if the source distribution is fixed. The same bound also applies to the entropy of the ciphertext. The bounds still hold if the secret message has been compressed before encryption.

This paper also illustrates that the lower bound only gives the minimum size of the pre-shared secret key. When a cipher system is used multiple times, this is no longer a reasonable measure for the portion of key consumed in each round. Instead, this paper proposes and justifies a new measure for key consumption rate. The existence of a fundamental tradeoff between the expected key consumption and the number of channel uses for conveying a ciphertext is shown. Optimal and nearly optimal secure codes are designed.

Index Terms

Shannon theory, information-theoretic security, perfect secrecy, joint source-encryption coding, one-time pad.

The material in this paper was presented in part at the IEEE International Symposium on Information Theory, St. Petersburg, July, 2011. The work of S.-W. Ho was supported by the Australian Research Council under an Australian Postdoctoral Fellowship as part of Discovery Project DP1094571. The work of T. H. Chan and A. Grant was also supported in part by ARC Discovery Project DP1094571.

I. INTRODUCTION

Cipher systems with *perfect secrecy* were studied by Shannon in his seminal paper [1] (see also [2]). With reference to Figure 1, a cipher system is defined by three components: a source message U , a ciphertext X and a key R . The key is secret common randomness shared by the sender and the legitimate receiver. The sender encrypts the message U , together with the key R , into the ciphertext X . This ciphertext will be transmitted to the legitimate receiver via a public channel. A cipher system is *perfectly secure*, or equivalently, satisfies a perfect secrecy constraint if the message U and the ciphertext X are statistically independent, $I(U; X) = 0$. In this case, an adversary who eavesdrops on the public channel and learns X (but does not have R) will not be able to infer any information about the message U . On the other hand, the legitimate receiver decrypts the message U from the received ciphertext X together with the secret key R . A cipher system is *error-free* (i.e., the probability of decoding error is zero) if $H(U | XR) = 0$.

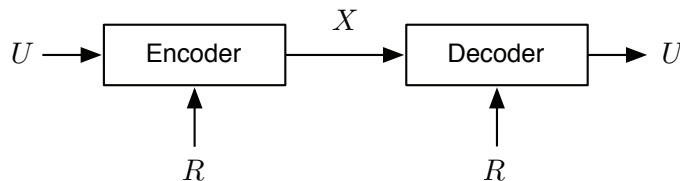


Fig. 1. A cipher system.

By considering a *deterministic cipher*, where X is a deterministic function of R and U , Shannon showed that the number of messages is equal to the number of possible ciphertexts, and that the number of different keys is not less than the number of messages [1, p. 681],

$$|\mathcal{X}| = |\mathcal{U}| \leq |\mathcal{R}|,$$

where \mathcal{X} , \mathcal{U} and \mathcal{R} are the respective supports of X , U and R . In order to design a perfectly secure cipher system protecting a source with *unknown source distribution* P_U , Shannon argued that

$$H(R) \geq \log |\mathcal{U}| \geq H(U). \quad (1)$$

He also made an important observation [1, p. 682] that

“the amount of uncertainty we can introduce into the solution cannot be greater than the key uncertainty”

In other words,

$$H(R) \geq H(U). \quad (2)$$

Massey [2] called (2) Shannon’s fundamental bound for perfect secrecy, and gave an information theoretic proof for this result. It is important to note that Massey’s proof [2] does not require U and R to be statistically independent.

Now, suppose U and R are indeed independent (which is common in practice). Our first main result, Theorem 1 improves (2), showing that for any source distribution P_U ,

$$P_R(r) \leq |\mathcal{U}|^{-1}, \quad \forall r. \quad (3)$$

As a consequence, we prove that for any cipher achieving perfect secrecy, the logarithm of the message sample size cannot be larger than the entropy of the secret key,

$$H(R) \geq \log |\mathcal{U}|. \quad (4)$$

Comparing with the first inequality in (1), we see that (4) is valid even if the source distribution P_U is fixed and known.

This paper is based on the model in Fig. 1. Despite its apparent simplicity, this is the most general encoder possible, and covers many interesting special cases. For example, suppose the distribution of U is non-uniform. One may expect that the optimal encoder will operate according to Fig. 2, by first compressing U and then encrypting the compressed output.

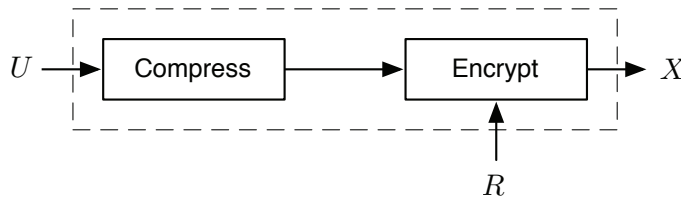


Fig. 2. Compression before encryption.

Roughly speaking, compression converts the source into a sequence of independent and identically distributed (i.i.d.) symbols. Theoretically, this can maximize the adversary’s decoding

error probability in some systems [3, Theorem 3]. Practically, the compressed output has a smaller file size and hence seem to require less key for encryption. This approach of compression before encryption was also proposed by Shannon [1, p. 682]. In fact, Shannon believed that, after removing redundancy in the source,

“a bit of key completely conceals a bit of message information”.

However a separated compression before encryption model is a special case of our more general model in Fig. 1. To certain extent, our model can be viewed as joint compression-encryption coding. Naturally, our results also apply to models such as Fig. 2, for which we will later prove

$$H(X) \geq \log |\mathcal{U}|.$$

This result, together with (4), in fact suggest that compression before encryption may not be useful if both perfect secrecy and error-free decoding are required.

Another major contribution of this paper is the introduction of a new concept of *expected key consumption* $I(R; UX)$. Previously in the literature, the amount of key required in a cipher system has been measured by the entropy of the common secret key. We will argue in this paper that $H(R)$ is only valid for measuring the *initial key requirement*, by which we mean the amount of secret randomness that must be shared between the sender and the legitimate receiver, prior to transmission of the ciphertext. Instead, key consumption should be measured by $I(R; UX)$. This new measure offers more insights, and in the second part of this paper, we will design efficient cipher system that can be used multiple times, where $I(R; UX)$ is one of the system parameters to be optimised.

Besides expected key consumption, we also want to minimize the *number of channel uses* required to transmit the ciphertext X from the source to the legitimate receiver. Naturally, we can encode the ciphertext X using a Huffman code [4]. Let $\lambda(X)$ be the codeword length. In this case, the expected codeword length $\mathbf{E}[\lambda(X)]$ satisfies $H(X) \leq \mathbf{E}[\lambda(X)] \leq H(X) + 1$. Note that for two random variables X and X' , it is possible that $H(X) < H(X')$, but $\mathbf{E}[\lambda(X)] > \mathbf{E}[\lambda(X')]$. One example is when $P_X = (0.3, 0.23, 0.2, 0.17, 0.1)$ and $P_{X'} = (0.25, 0.25, 0.25, 0.15, 0.1)$. However, we still use $H(X)$ instead of $\mathbf{E}[\lambda(X)]$ as a measure for the number of channel uses required in a cipher system for two reasons: first, $H(X)$ is a lower bound for $\mathbf{E}[\lambda(X)]$ and in fact a very good estimate for $\mathbf{E}[\lambda(X)]$; second, the problem itself is more tractable when using $H(X)$, instead of $\mathbf{E}[\lambda(X)]$.

We will show that there exists a fundamental tradeoff between the expected key consumption and the number of channel uses. In fact, if the source distribution is not uniform, then the minimum expected key consumption and the minimum number of channel uses cannot be simultaneously achieved. We will also show that code design achieving minimum expected key consumption depends on whether the source distribution P_U has irrational probability masses or not. Optimal code will be proposed for P_U which has only rational probability masses.

Organization: In Section II, we consider one-shot systems, where there is a single message to be securely transmitted. We formalize the system model, and new bounds on $H(R)$ and $H(X)$ will be derived. In Section III, we will consider the case where cipher system is used multiple times. New system parameters including $I(R; UX)$ will be defined and justified. Section IV will focus on two regimes corresponding to minimal expected key consumption and minimal number of channel uses. The existence of a fundamental non-trivial tradeoff will be illustrated. In Section V, the performance of compression-before-encryption will be evaluated.

Notation. Random variables are denoted by capital letters, e.g. X , and their particular realizations are denoted by small letters, x . Supports of random variables are denoted by calligraphic letters, \mathcal{X} .

II. KEY REQUIREMENTS FOR ONE-SHOT CIPHERS

Definition 1 (Error free perfect secrecy system): A cipher system (R, U, X) is called an *Error-free Perfect-Secrecy (EPS)* system if

$$I(U; X) = 0, \tag{5}$$

$$H(U | RX) = 0, \tag{6}$$

$$I(U; R) = 0. \tag{7}$$

Here, (5) ensures perfect secrecy, via independence of the ciphertext X and source message U . An eavesdropper learning X can infer no information about the message U . The constraint (6) ensures that the receiver can reconstruct U from R and X without error. Finally (7) requires that the shared secret key R is independent of the message U .

The constraints (5) and (6) were originally used in [2] to prove Shannon's fundamental bound (2) for perfect secrecy. The only additional constraint in Definition 1 is (7). In practice, R is usually shared prior to the independent generation of the message U . This is a strong practical

motivation for (7). Furthermore, Definition 1 admits the general case of probabilistic encoding. For the receiver, it is however sufficient to consider deterministic decoding since by (6), U is a function of R and X . In other words, there exists a decoding function g such that

$$P_{URX}(u, r, x) = P_{RX}(r, x) \mathbf{1}\{u = g(r, x)\}. \quad (8)$$

Theorem 1 (Lower bounds on $H(X)$ and $H(R)$): Let (R, U, X) be an error free perfect secrecy system, satisfying (5) – (7) according to Definition 1, and suppose P_U is known. Then

$$\max_{x \in \mathcal{X}} P_X(x) \leq |\mathcal{U}|^{-1}, \quad (9)$$

and

$$\max_{r \in \mathcal{R}} P_R(r) \leq |\mathcal{U}|^{-1}, \quad (10)$$

where \mathcal{U} is the support of the message U . Consequently,

$$\log |\mathcal{U}| \leq H(X), \quad (11)$$

with equality if and only if $P_X(x) = |\mathcal{U}|^{-1}$ for all $x \in \mathcal{X}$. Also,

$$\log |\mathcal{U}| \leq H(R), \quad (12)$$

with equality if and only if $P_R(r) = |\mathcal{U}|^{-1}$ for all $r \in \mathcal{R}$. If the source distribution is not uniform, $H(X)$ and $H(R)$ are strictly greater than $H(U)$.

Proof: For any $x \in \mathcal{X}$,

$$|\mathcal{U}|P_X(x) = \sum_u P_X(x) \quad (13)$$

$$= \sum_u P_{X|U}(x|u) \quad (14)$$

$$= \sum_u \sum_{r:P_{URX}(u,r,x)>0} \frac{P_{URX}(u,r,x)}{P_U(u)} \quad (15)$$

$$= \sum_u \sum_{r:P_{URX}(u,r,x)>0} \frac{P_{RX}(r,x) \mathbf{1}\{u=g(r,x)\}}{P_U(u)} \quad (16)$$

$$= \sum_{r:P_{RX}(r,x)>0} \frac{P_{RX}(r,x)}{P_U(g(r,x))} \quad (17)$$

$$= \sum_{r:P_{RX}(r,x)>0} P_{RX}(r,x) \frac{P_{X|UR}(x|g(r,x),r) P_R(r)}{P_{URX}(g(r,x),r,x)} \quad (18)$$

$$= \sum_{r:P_{RX}(r,x)>0} P_{X|UR}(x|g(r,x),r) P_R(r) \quad (19)$$

$$\leq \sum_{r:P_{RX}(r,x)>0} P_R(r) \quad (20)$$

$$\leq 1, \quad (21)$$

where (14), (16), (18) and (21) follow from (5), (8), (7) and (8), respectively. This establishes (9).

Let P_B be a uniform distribution with support \mathcal{U} . Since P_X is always majorized¹ by P_B from (9), [7, Theorem 10] shows that

$$H(P_X) \geq H(P_B) + D(P_B||P_X) \quad (22)$$

$$\geq H(P_B) \quad (23)$$

$$= \log |\mathcal{U}|, \quad (24)$$

and hence (11) is verified. Note that [7, Theorem 10] can still be applied even if X may be defined on a countably infinite alphabet. If $H(P_X) = \log |\mathcal{U}|$, equality in (23) holds so that $P_X \equiv P_B$. Finally, (10) and (12) follow from the symmetric roles of X and R in (5) – (7). ■

¹A good introduction to majorization theory can be found in [6]. In this proof, we just need the definition of “majorized by” which can also be found in [7, Definition 1]

Corollary 2: No error free perfect secrecy system can be constructed if the source message U has a countably infinite support or a support with unbounded size.

Proof: Assume in contradiction that an EPS system exists for a source message $U \sim P_U$ with countably infinite support, $|\mathcal{U}| = \infty$. Note that (13) – (21) are still valid in this case. However, the conclusion that $|\mathcal{U}|P_X(x) \leq 1$ for any $x \in \mathcal{X}$ contradicts $|\mathcal{U}| = \infty$. ■

The following three remarks emphasize some of the (perhaps unexpected) consequences of Theorem 1.

- 1) One could naturally expect that $H(U)$ is the critical quantity setting a lower bound on $H(R)$ and $H(X)$. However, Theorem 1 shows that $H(R)$ and $H(X)$ can be arbitrarily large, as long as the size of the support of U is also arbitrarily large, *even when $H(U)$ is small*.
- 2) One may further expect that $\log |\mathcal{U}| \leq H(R)$ is tight only if the source distribution P_U is unknown. However, (11) and (12) show that fixing P_U does not reduce the lower bounds on either the initial key requirement, or the number of channel uses required to convey the ciphertext.
- 3) If the source message U is defined on a countably infinite alphabet, it is not possible to design an error free perfect secrecy system (Corollary 2). Therefore, if a cipher system is required for such a source, at least one of the constraints (5) – (7) must be relaxed.

The following example compares Shannon's fundamental bound (2) with Theorem 1. It also illustrates that the quantity $H(R)$ is insufficient for determination of the requirements on the secret key R .

Example 1: Suppose $P_U = (0.3, 0.3, 0.3, 0.1)$ so that $H(U) = 1.895$ bits and $\log |\mathcal{U}| = 2$ bits.

- 1) Consider R chosen independently of U according to $P_R = (0.4, 0.2, 0.2, 0.2)$ so that $H(R) = 1.922$ bits and $H(U) < H(R) < \log |\mathcal{U}|$. Although P_R satisfies Shannon's fundamental bound (2), Theorem 1, in particular (12), shows this choice of key R is insufficient to achieve error free perfect secrecy.
- 2) Consider $P_R = (0.4, 0.15, 0.15, 0.15, 0.15)$ so that $H(R) = 2.171$ bits and $H(U) < \log |\mathcal{U}| < H(R)$. However, this choice of key R is insufficient for error free perfect secrecy, since from (10), $\max_r P_R(r) = 0.4 > 0.25 = |\mathcal{U}|^{-1}$.

Theorem 1 not only applies to systems of the form shown in Fig. 1 (which includes Fig. 2 as

a special case), but also to multi-letter variations. For example, we can accumulate n symbols from the source (U_1, U_2, \dots, U_n) and treat these n symbols together as one super-symbol U . It is reasonable to consider finite n because practical systems have only finite resources to store the super-symbol. Unless the source has some special structure, the distribution of U cannot be uniform for any n if the U_i are not uniform. For example, if the source is stationary and memoryless, accumulating symbols will only make $H(X)$ and $H(R)$ grow with $n \log |\mathcal{U}|$.

One may argue that the coding rate of $H(X)$ could be reduced because the sender and receiver share the same side information R and $I(X; R) > 0$ is possible. In other words, a compressor may be appended to the encoder in Fig. 1 in order to reduce the size of the ciphertext. This configuration is shown in Fig. 3. However, we cannot simply apply the results from source coding with side information here, because the ciphertext still needs to satisfy the security constraint. If the new output Y satisfies the perfect secrecy and zero-error constraints, $I(U; Y) = H(U|RY) = 0$, then (R, U, Y) in Fig. 3 is simply another EPS system, governed by Theorem 1.

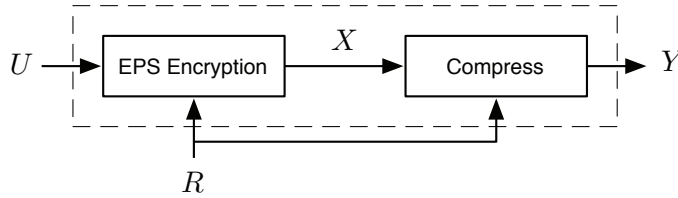


Fig. 3. Compressing the output of an EPS cipher.

To complete this section, we show that the lower bounds (11) and (12) are simultaneously achievable using a one-time pad [8].

Definition 2 (One-time pad): Without loss of generality, let $\mathcal{U} = \{0, \dots, M-1\}$ be the support of U . Let R be independent of U and uniformly distributed in \mathcal{U} and let X be generated according to the *one-time pad* as $X = (U + R) \bmod M$. Then U can be recovered via $(X + R) \bmod M$.

It is easy to verify that (5) – (7) are satisfied and $H(X) = H(R) = \log M$. Therefore, we have proved the following theorem.

Theorem 3 (Achieving the minimum $H(X)$ and $H(R)$): Let \mathcal{U} be the support of U . The one-time pad of Definition 2 is an EPS system achieving $H(X) = \log |\mathcal{U}|$ and $H(R) = \log |\mathcal{U}|$.

III. MULTIPLE MESSAGES AND KEY CONSUMPTION

In Section II, Theorem 3 proved that the one-time pad is “optimal” in the sense that it simultaneously minimizes $H(X)$ and $H(R)$. This immediately suggests that the one-time pad leaves no room for improvement. However, this conclusion in fact stems from a folk theorem that the “required size of the secret key” is measured by the key entropy. The hidden assumption behind this folklore is that the *cipher system is used only once*. In typical practice, a cipher system will be used repeatedly for the transmission of multiple messages.

Consider the following scenario. Suppose an initial secret key R is delivered to the sender and the receiver prior to commencement of message transmission. Now, suppose the sender uses this key to encrypt a message U , which is then delivered to the receiver over the public channel. Clearly, some portion of the secret randomness R has now been used. The central question is as follows: Can the sender and receiver continue to securely communicate without first receiving a new key? For example, if U is a single bit and R is a 100-bit random key, it is indeed likely that another message can be securely transmitted. The natural questions are: What is the maximum size of the second message? Alternatively, how much of the key R was consumed in the first round of transmission? Below, we will show that when an error free perfect secrecy system is used multiple times, the key consumption should not be measured by $H(R)$ but by $I(R; UX)$. In fact, with respect to our definitions, we will exhibit systems with key consumption that can be made arbitrarily close to $H(U)$.

The following example illustrates some of the basic ideas which will be elaborated in this section.

Example 2: Suppose the sender and the receiver share a secret key $R = \{B_1, B_2, \dots, B_n\}$, where all of the B_i , $i = 1, 2, \dots, n$ are independent and uniformly distributed over $\{0, 1\}$. Let $P_U(0) = 0.5$ and $P_U(1) = P_U(2) = 0.25$. Construct a new random variable U' such that

$$U' = \begin{cases} (0, B_{n+1}), & U = 0 \\ (1, 0), & U = 1 \\ (1, 1), & U = 2 \end{cases} \quad (25)$$

where B_{n+1} is generated by the sender independently of U and R such that $P_{B_{n+1}}(0) = P_{B_{n+1}}(1) = 0.5$.

Let $K = (B_1, B_2)$ and $X = U' \oplus K$. Upon receiving X , the receiver can decode U from X and K , where K is solely a function of R . In fact, if $U = 0$, the receiver can further decode B_{n+1} . Let

$$R' = \begin{cases} (B_3, B_4, \dots, B_n), & U \in \{1, 2\} \\ (B_3, B_4, \dots, B_n, B_{n+1}), & U = 0. \end{cases}$$

We refer to R' as the *residual secret randomness* shared by the sender and the receiver. Note that R' may not be a deterministic function of R , as the new shared common randomness can be generated by a probabilistic encoder. According to (25), a new random bit is secretly transmitted from the sender to the receiver when $U = 0$. After the system is used once, the expected key consumption is therefore given by

$$P_U(0) \cdot 1 + P_U(1) \cdot 2 + P_U(2) \cdot 2 = 1.5 = H(U), \quad (26)$$

which happens to also equal $I(R; UX)$. It turns out that this is not mere coincidence.

We now define three parameters whose operational meanings are justified in the rest of this section.

Definition 3: The *residual secret randomness* of an error free perfect secrecy system is

$$H(R | UX).$$

Definition 4: The *expected key consumption* of an error free perfect secrecy system is

$$I(R; UX).$$

Definition 5: The *excess key consumption* of an error free perfect secrecy system is

$$I(R; X).$$

Roughly speaking, we will show that after an EPS system is used once, $H(R | UX)$ is the amount of remaining key that can be used for encryption of the next message. Since the sender and the receiver initially share a quantity $H(R)$ of secret randomness, the key consumption is equal to $H(R) - H(R|UX) = I(R; UX)$. We will provide achievable schemes to show that the minimal key consumption is $H(U)$ and hence, the excess key consumption is $I(R; UX) - H(U)$ which is equal to $I(R; X)$ in an EPS system.

We first justify Definition 3. Consider the scenario of Fig. 4 in which the sender and receiver share a secret key R , and two EPS systems are used sequentially by the sender to securely

transmit two (possibly correlated) messages U and V . In the first round, the sender encodes the message U into X , which is transmitted to the receiver as described in Section II. In the second round, the sender further encodes V (or more generally both U and V) into Y , which is then transmitted to the receiver. As before, we require $H(U | RX) = H(V | RXY) = 0$ and $I(UV; XY) = 0$ to ensure zero-error decoding and perfect secrecy.

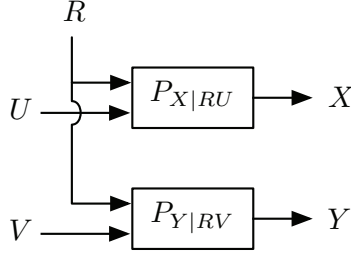


Fig. 4. Using an error free perfect secrecy system twice.

Theorem 4 (Justification 1): Consider the two-round error free perfect secrecy system of Fig. 4. If

$$I(UV; XY) = H(U | RX) = H(V | RXY) = 0, \quad (27)$$

then the entropy of the second message V conditioning on the first message U is upper bounded by the residual secret randomness,

$$H(V | U) \leq H(R | UX). \quad (28)$$

Proof: Note that

$$\begin{aligned} & H(R | UX) - H(V | U) + I(UV; XY) + H(U | RX) + H(V | RXY) \\ &= I(VY; U | RX) + I(RU; Y | X) + I(U; X) + H(U | RXY) + H(R | UVXY) \geq 0. \end{aligned}$$

Together with (27), (28) is verified. ■

Theorem 4 implies that the maximum amount of information which can be secretly transmitted in the second round is upper bounded by the residual secret randomness $H(R | UX)$, suggesting that $H(R | UX)$ is indeed measures the amount of key unused in the first round. Equivalently, the amount of key that has been consumed in the first round is equal to

$$I(R; UX) = H(R) - H(R | UX).$$

Whereas Theorem 4 justifies the residual key $H(R | UX)$ as bounding the entropy of the second round message, we now offer an alternative justification, showing that the size of the key that can be extracted after n uses of an EPS system is about $nH(R | UX)$.

Consider generation of a new secret key as shown in Fig. 5. Suppose a sequence of EPS systems $\{(U_i, R_i, X_i)\}_{i=1}^n$ has been used by a sender and a receiver where (U_i, R_i, X_i) are i.i.d. with generic distribution P_{URX} . We use (U, R, X) to denote the generic random variables. In order to securely send additional messages, the sender and the receiver aim to establish a new secret key $S^m = (S_1, \dots, S_m)$, where the S_i are i.i.d. with generic distribution P_S . To generate the new key S^m , we assume that the sender can send a secret message A to the receiver. The new secret key S^m will be used to encrypt a second sequence of messages V^m , generating a ciphertext sequence Y^m such that $\{(V_i, S_i, Y_i)\}_{i=1}^m$ is another sequence of EPS systems.

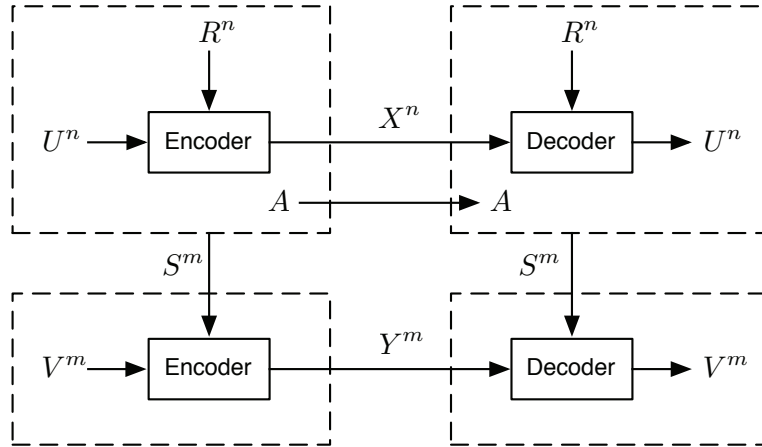


Fig. 5. Generating a new secret key S^m .

Assume

$$I(V^m; S^m U^n X^n) = 0 \quad (29)$$

$$I(U^n X^n; Y^m | V^m S^m) = 0 \quad (30)$$

$$H(S^m | R^n U^n X^n A) = 0 \quad (31)$$

$$I(S^m; U^n X^n) = 0. \quad (32)$$

These assumptions adopted with the following reasoning. We assume in (29) that the new message V^m is generated independently of the previous uses of the EPS systems. Also, (30)

holds due to $(U^n, X^n) - (V^m, S^m) - Y^m$ forms a Markov chain. The sender and the receiver can agree on S^m without error due to (31). The justification of (32) is given as follows.

Although $\{(U_i, R_i, X_i)\}_{i=1}^n$ and $\{(V_i, S_i, Y_i)\}_{i=1}^m$ are individually sequences of EPS systems, it is possible that their combination is not secure, $I(X^n, Y^m; U^n, V^m) > 0$. For example, suppose U^n and V^m are i.i.d. with uniform distribution and $m = n$. If $S^m = U^n$, then using a one-time pad, $I(V^m; Y^m) = 0$ but $I(X^n, Y^m; U^n, V^m) \geq H(U^n)$. The following theorem shows that joint EPS systems satisfying (29) – (32) are still perfectly secure.

Theorem 5: Consider two sequences of i.i.d. EPS systems $\{(U_i, R_i, X_i)\}_{i=1}^n$ and $\{(V_i, S_i, Y_i)\}_{i=1}^m$ satisfying (29) – (32). Then the joint EPS system is still perfectly secure,

$$I(X^n Y^m; U^n V^m) = 0. \quad (33)$$

Proof: By assumption

$$\begin{aligned} I(U^n; X^n) &= I(V^m; Y^m) = I(V^m; S^m, U^n, X^n) = \\ I(U^n, X^n; Y^m | V^m, S^m) &= I(S^m; U^n, X^n) = 0. \end{aligned} \quad (34)$$

Note that

$$\begin{aligned} I(S^m; U^n, X^n) &+ I(V^m; S^m, U^n, X^n) + I(V^m; Y^m) + \\ I(U^n; X^n) &+ I(U^n, X^n; Y^m | V^m S^m) - I(X^n, Y^m; U^n, V^m) \\ &= I(U^n, X^n; S^m | V^m, Y^m) + I(V^m; S^m) + I(X^n; Y^m) + I(U^n; V^m) \geq 0. \end{aligned}$$

Together with (34), $I(X^n, Y^m; U^n, V^m) \leq 0$. Since $I(X^n, Y^m; U^n, V^m) \geq 0$, (33) is verified. ■

In order to generate a new key S^m , a secret auxiliary random variable A is sent from the sender to the receiver. Here, A is generated by a probabilistic encoder with $\{(R_i, U_i, X_i)\}_{i=1}^n$ as input. In Example 2 above, suppose we wanted to restore n secret bits after the system is used once. Then A is a fair bit if $U = 0$ and A consists of two fair bits if $U = 1$ or 2 . We measure the expected size of A by $H(A | R^n U^n X^n)$. Since we can directly treat A as the new secret key S^m , it is reasonable to expect that $H(S^m) \geq H(A | R^n U^n X^n)$. Therefore, it is of interest to know by how much $H(S^m)$ can exceed $H(A | R^n, U^n, X^n)$ for a given sequence of EPS systems. The following theorem shows that the secret randomness, which can be extracted from $\{(R_i, U_i, X_i)\}_{i=1}^n$ with help from A , is measured by the residual secret randomness $H(R | U, X)$.

Theorem 6 (Justification 2): Consider two sequences of i.i.d. EPS systems $\{(U_i, R_i, X_i)\}_{i=1}^n$ and $\{(V_i, S_i, Y_i)\}_{i=1}^m$ and any A . If (29) – (32) are satisfied, then

$$H(S^m) - H(A | R^n U^n X^n) \leq nH(R | UX). \quad (35)$$

On the other hand, it is possible to generate S^m such that (29) – (32) are satisfied and

$$H(S^m) - H(A | R^n U^n X^n) \geq nH(R | UX) - \log 2 \quad (36)$$

for a sufficiently large m such that

$$\max_{s^m} P_{S^m}(s^m) < \min_{r^n, u^n, x^n} P_{R^n | U^n X^n}(r^n | u^n, x^n).$$

Proof: We first prove (35) by showing that

$$H(S^m) = I(S^m; U^n X^n) + H(S^m | AR^n U^n X^n) + I(S^m; AR^n | U^n X^n) \quad (37)$$

$$= I(S^m; AR^n | U^n X^n) \quad (38)$$

$$\leq H(AR^n | U^n X^n) \quad (39)$$

$$= H(A | R^n U^n X^n) + H(R^n | U^n X^n) \quad (40)$$

$$= H(A | R^n U^n X^n) + nH(R | UX), \quad (41)$$

where (38) follows from (31) – (32) and (41) follows from the fact that $\{(U_i, R_i, X_i)\}_{i=1}^n$ is a sequence of i.i.d. EPS systems.

The proof of the achievability part in (36) is via construction. With reference to Fig. 6, consider two partitions of the unit interval into disjoint “cells”. The width of cell i in the first partition is $P_{S^m}(i)$ for $1 \leq i \leq |\mathcal{S}|^m$, where \mathcal{S} is the support of S_i . Consider $U^n = u^n$ and $X^n = x^n$. The width of cell i in the second partition is $P_{R^n | U^n, X^n}(i | u^n, x^n)$ for $1 \leq i \leq |\mathcal{R}|^n$. The distribution of A is constructed to divide the second partition as shown in Fig. 6. To simplify notations, we consider the support of R^n to be a set of consecutive integers $\{1, \dots, |\mathcal{R}'|\}$ when $U^n = u^n$ and $X^n = x^n$. Suppose $R^n = r$ and let

$$b = \max \left\{ j : \sum_{i=1}^j P_S(i) > \sum_{i=1}^{r-1} P_{R^n | U^n, X^n}(i | u^n, x^n) \right\}. \quad (42)$$

For $j \geq 1$, A is defined by $P_A(j) = \frac{a(j)}{P_{R^n | U^n, X^n}(r | u^n, x^n)}$, where

$$\sum_{i=1}^j a(i) = \min \left\{ \sum_{i=1}^{b+j-1} P_S(i), \sum_{i=1}^r P_{R^n | U^n, X^n}(i | u^n, x^n) \right\} - \sum_{i=1}^{r-1} P_{R^n | U^n, X^n}(i | u^n, x^n). \quad (43)$$

1	2	3	4	5	6	...	$ \mathcal{S} ^m$	
1		2	3		4		...	$ \mathcal{R}' $
1	2	1	2	1	2	3	1	2

Fig. 6. An assignment of A .

For the example in Fig. 6, when $R^n = 1$,

$$P_A(1) = \frac{P_{S^m}(1)}{P_{R^n|U^n, X^n}(1 | u^n, x^n)} = 1 - P_A(2). \quad (44)$$

When $R^n = 2$,

$$P_A(1) = \frac{P_{S^m}(1) + P_{S^m}(2) - P_{R^n|U^n, X^n}(1 | u^n, x^n)}{P_{R^n|U^n, X^n}(2 | u^n, x^n)} = 1 - P_A(2). \quad (45)$$

By definition S^m is determined from R^n and A for any fixed $U^n = u^n$ and $X^n = x^n$. On the other hand, A is also determined from S^m and R^n . Therefore,

$$H(S^m | A, R^n, U^n, X^n) = H(A | S^m, R^n, U^n, X^n) = 0. \quad (46)$$

By choosing m sufficiently large, such that

$$\max_{s^m} P_{S^m}(s^m) < \min_{r^n, u^n, x^n} P_{R^n|U^n, X^n}(r^n | u^n, x^n), \quad (47)$$

R^n can take at most two possible values for any given (S^m, U^n, X^n) and hence

$$H(R^n | S^m, U^n, X^n) \leq \log 2. \quad (48)$$

Therefore,

$$H(A | R^n U^n X^n) \quad (49)$$

$$= I(A; S^m | R^n U^n X^n) + H(A | S^m R^n U^n X^n) \quad (50)$$

$$= I(A; S^m | R^n U^n X^n) + H(S^m | A R^n U^n X^n) \quad (51)$$

$$= H(S^m | R^n U^n X^n) \quad (52)$$

$$= H(S^m) - H(R^n | U^n X^n) + H(R^n | S^m U^n X^n) - I(S^m; U^n, X^n) \quad (53)$$

$$\leq H(S^m) - H(R^n | U^n X^n) + H(R^n | S^m U^n X^n) \quad (54)$$

$$\leq H(S^m) - H(R^n | U^n X^n) + \log 2, \quad (55)$$

where (51) and (55) follow from (46) and (48), respectively. Since $\{(U_i, R_i, X_i)\}_{i=1}^n$ is a sequence of i.i.d. EPS systems, (36) is verified.

For any (U^n, X^n) , the same $P_{S^m|U^n, X^n} \equiv P_{S^m}$ is generated. Therefore, (32) is verified. Since S^m is determined by (R^n, U^n, X^n, A) , (31) is verified, and (29) can also be verified as V^m is independent of (R^n, U^n, X^n, A) . Finally, (30) is due to the fact that $\{(V_i, S_i, Y_i)\}_{i=1}^m$ is a sequence of EPS systems. ■

Roughly speaking, Theorem 6 shows that for large n and m , the optimal algorithm with the help of A can extract approximately

$$nH(R | UX)$$

bits of residual secret randomness from $\{(R_i, U_i, X_i)\}_{i=1}^n$. In [9], we considered another algorithm generating a new secret key with asymptotic rate $H(R | UX)$ without using an auxiliary secret random variable. As the sender and receiver initially share $nH(R)$ bits of secret randomness, the expected key consumption for each use of the EPS system is

$$H(R) - H(R | UX) = I(R; UX),$$

the quantity proposed in Definition 4. Next, we exhibit an important property of $I(R; UX)$.

Theorem 7: In an error free perfect secrecy system, the expected key consumption is lower bounded by the source entropy,

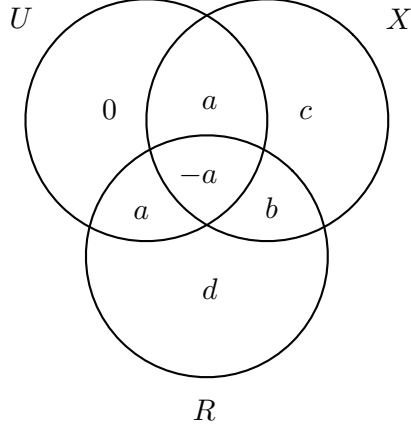
$$I(R; UX) \geq H(U), \quad (56)$$

where equality holds if and only if $I(R; X) = 0$.

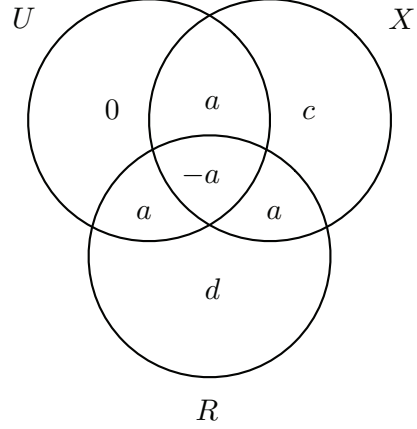
Proof: The information diagram for the random variables U, X, R involved in an error free perfect secrecy system satisfying (5) – (7) is shown in Fig. 7(a). It is easy to verify that

$$I(X; R) = I(R; UX) - H(U). \quad (57)$$

Since $I(X; R) \geq 0$, Theorem 7 is proved. ■



(a) General EPS system



(b) Minimum expected key consumption, achieving equality in (56)

Fig. 7. Information diagrams.

In Section IV-A, we will describe several EPS coding schemes achieving $I(R; UX) = H(U)$. Therefore, $I(X; R)$ measures the difference between the expected key consumption of an EPS system and the minimum possible key consumption, again justifying Definition 5. The information diagram for the optimal case $I(X; R) = 0$ is shown in Fig. 7(b).

We summarize this section in the following three remarks.

- 1) Theorems 4 and 6 provide strong justification of $I(R; UX)$ as the expected key consumption required to achieve error free perfect secrecy. Theorem 7 shows that the expected key consumption cannot be less than the source entropy. Recall that Theorem 1 gives the lower bound on the initial key requirement. Therefore, we have distinguished between two different concepts (a) expected key consumption in a multi-round system and (b) the initial key requirement for a one-shot system. In contrast to the bound $H(R) \geq H(U)$ [1],

[2], Theorem 7 more precisely describes the role of $H(U)$ in an error free perfect secrecy system.

2) From (5)–(7) we can show that

$$H(R) = H(U) + I(X; R) + H(R | UX). \quad (58)$$

Thus the key entropy $H(R)$ consists of three parts: the randomness used to protect the source, the excess key consumption and the residual secret randomness.

3) If the source distribution is uniform, Example 3 below shows that the one-time pad achieves minimal key consumption.

Example 3 (Uniform source distribution): Suppose U and R are independent and are uniformly distributed on the sets $\{0, 1, \dots, 2^i - 1\}$ and $\{0, 1, \dots, 2^j - 1\}$, respectively, where $i \leq j$. In order to derive a coding system satisfying (5) – (7), we can first extract i random bits R' from R and construct X as the modulo-two addition of the binary representation of U and R' . Then

$$I(R; UX) = H(R) - H(R | UX) \quad (59)$$

$$= H(R) - H(R | R') \quad (60)$$

$$= j - (j - i) \quad (61)$$

$$= H(U). \quad (62)$$

IV. TRADEOFF BETWEEN KEY CONSUMPTION AND NUMBER OF CHANNEL USES

Example 3 shows that the one-time pad simultaneously achieves the minimal expected key consumption and the minimum number of channel uses for a uniform source. However for general non-uniform sources, we will show that there is a non-trivial tradeoff between these two quantities.

We will consider two important regimes. First, in Section IV-A, we will consider the regime in which ciphers minimize the key consumption $I(R; UX)$. Conversely, in Section IV-B we consider systems which minimize the number of channel uses $H(X)$.

We shall demonstrate the existence of a fundamental, non-trivial tradeoff between the expected key consumption and the number of channel uses. Our main results, Theorem 1 proved earlier, and Theorems 7 – 14 to be proved below, are summarized in Fig. 8.

Point 1 is due to Theorem 14 in Section IV-B below, and has the smallest $I(R; UX)$ among all EPS systems with $H(X) = \log |\mathcal{U}|$. We shall show that this point can always be achieved by one-time pad.

Point 2 has the smallest $H(X)$ among all the EPS systems with $I(R; UX) = H(U)$. For this point, Theorem 11 in Section IV-A gives the lower bound on $H(X)$ which is strictly greater than $\log |\mathcal{U}|$ if P_U is not uniform.

If P_U has only rational probability masses, Theorem 8 in Section IV-A below shows that Point 3 can be achieved by a generalization of the one-time pad, the *partition code* (to be introduced in Definition 6).

If all the probabilities masses in P_U are the integer multiples of the smallest probability mass in P_U , then Point 2 coincides with Point 3 by the partition code shown in Theorem 13. Otherwise, Point 3 can differ from Point 2 which will be demonstrated in Example 4.

The existence, continuity and non-increasing in $H(X)$ properties of the curved portion of the tradeoff curve are established in Section IV-C.

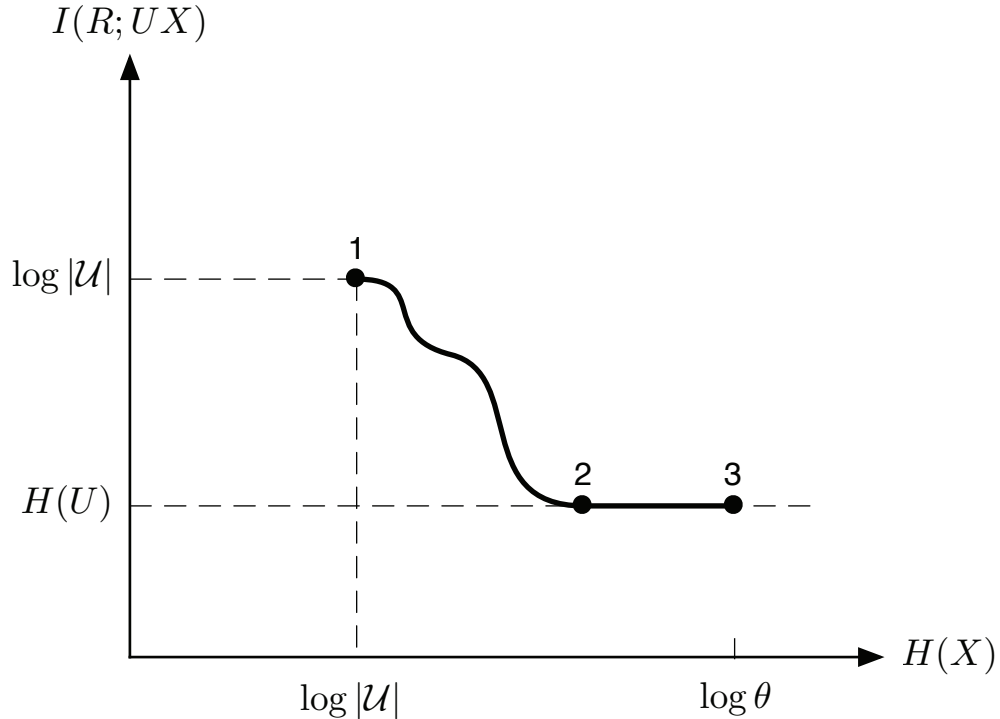


Fig. 8. Tradeoff between $I(R; UX)$ and $H(X)$.

A. Minimal expected key consumption

We first consider EPS systems which achieve minimal expected key consumption. From Theorem 7, an error free perfect secrecy system with minimal key consumption satisfies (5)–(7) and

$$I(X; R) = 0. \quad (63)$$

We now generalize the one-time pad to achieve minimal key consumption for source distributions containing only rational probability masses.

Definition 6 (Partition Code $\mathcal{C}(\Psi)$): Assume that U is a random variable defined on $\{1, \dots, \ell\}$. Let $\Psi = (\psi_1, \psi_2, \dots, \psi_\ell)$ and let $\theta = \sum_{i=1}^{\ell} \psi_i$ where ψ_i and θ are positive integers. Let A' be a random variable such that

$$\Pr(A' = j \mid U = i) = \begin{cases} \frac{1}{\psi_i} & \text{if } 1 \leq j \leq \psi_i, \\ 0 & \text{otherwise.} \end{cases}$$

Let $A = \sum_{i=1}^{U-1} \psi_i + A' - 1$, R be uniformly distributed on the set $\{0, 1, \dots, \theta - 1\}$ and $X = A + R \bmod \theta$. The so defined cipher system (R, U, X) is called the *partition code* $\mathcal{C}(\Psi)$.

Note that one-time pad is a special case of partition code when $\Psi = (1, 1, \dots, 1)$.

It can be proved directly that a partition code satisfies (5) – (7) and hence is an EPS system. Furthermore, we can verify that

$$H(X) = H(R) = \log \theta, \quad (64)$$

and

$$I(R; UX) = \sum_{i=1}^{\ell} P_U(i) \log \frac{\theta}{\psi_i}, \quad (65)$$

where (65) is from $H(X \mid U, R) = H(A \mid U, R) = \sum_{i=1}^{\ell} P_U(i) \log \psi_i$ together with (64).

Let Q_U be the probability distribution such that $Q_U(i) = \psi_i/\theta$. Then (65) can be rewritten as

$$I(R; UX) = H(U) + D(P_U \parallel Q_U), \quad (66)$$

where $D(\cdot \parallel \cdot)$ is the relative entropy [4]. Consequently, we have the following theorem.

Theorem 8: Suppose the probability mass $P_U(i)$ is rational for all $i = 1, \dots, \ell$. Let θ be an integer such that $\theta \cdot P_U(i)$ is also an integer for all i , and let $\Psi = (\psi_1, \psi_2, \dots, \psi_\ell)$ with

$\psi_i = \theta \cdot P_U(i)$. Then the EPS system (R, U, X) induced by the partition code $\mathcal{C}(\Psi)$ achieves the lower bound in (56), namely $I(R; UX) = H(U)$.

In the following theorem, we prove that if the source distribution P_U is not rational, then partition code will not achieve zero key-excess with finite X or R . Its proof is deferred to Section VI.

Theorem 9: Suppose U , X , and R satisfy (5) – (7) and (63). If there exists $u \in \mathcal{U}$ such that $P_U(u)$ is irrational, then the support of X and R cannot be finite.

Although it is difficult to construct codes satisfying (5) – (7) and (63) for P_U having irrational probability masses, Theorem 7 still gives a tight bound on $I(R; UX)$ as shown in the following theorem.

Theorem 10: Suppose the support of P_U is a finite set of integers $\{1, \dots, \ell\}$. Let $\Psi = (\psi_1, \dots, \psi_{\ell+1})$ with

$$\psi_i = \begin{cases} \lfloor P_U(i)\theta \rfloor, & 1 \leq i \leq \ell, \\ \theta - \sum_{i=1}^{\ell} \lfloor P_U(i)\theta \rfloor, & i = \ell + 1. \end{cases}$$

Assume that θ is large enough such that $\lfloor P_U(i)\theta \rfloor \geq 1$ for all $1 \leq i \leq \ell$. For the partition code $\mathcal{C}(\Psi)$, $I(R; UX) \rightarrow H(U)$ as $\theta \rightarrow \infty$.

Proof: Consider a probability distribution Q_U with $Q_U(i) = \psi_i/\theta$ for $1 \leq i \leq \ell + 1$. As $\theta \rightarrow \infty$, Q_U converges pointwise to P_U and hence $D(P_U \| Q_U) \rightarrow 0$ for finite ℓ . The theorem thus follows from (66). ■

In addition to minimizing the key consumption $I(R; UX)$, we may also want to simultaneously minimize $H(X)$, which is the number of channel uses required to convey the ciphertext X . The following theorem and corollary illustrate that the zero key-excess condition can be very harsh, requiring the EPS system to have a very large $H(R)$ and $H(X)$, even for very simple sources.

Theorem 11 (EPS systems with minimal $I(R; UX)$): Let \mathcal{X} , \mathcal{R} and \mathcal{U} be the respective supports of random variables X , R , and U satisfying (5)–(7) and (63). Then

$$\max_{x \in \mathcal{X}} P_X(x) \leq \min_{u \in \mathcal{U}} P_U(u) \quad (67)$$

and

$$\max_{r \in \mathcal{R}} P_R(r) \leq \min_{u \in \mathcal{U}} P_U(u). \quad (68)$$

Proof: Consider any $u \in \mathcal{U}$ and $x \in \mathcal{X}$. By definition, $P_U(u) > 0$ and $P_X(x) > 0$. From (5), we have $P_{UX}(u, x) = P_U(u)P_X(x) > 0$. Consequently, there exists $r \in \mathcal{R}$ such that $P_{UXR}(u, x, r) > 0$. Notice that

$$P_{UXR}(u, x, r) = P_{XR}(x, r) \quad (69)$$

$$= P_X(x)P_R(r), \quad (70)$$

where (69) is due to (6) and (70) is due to (63). On the other hand,

$$P_{UXR}(u, x, r) \leq P_{UR}(u, r) \quad (71)$$

$$= P_U(u)P_R(r), \quad (72)$$

where (72) is due to (7). Finally, as $P_R(r) > 0$, we have $P_X(x) \leq P_U(u)$ and (67) follows. Due to the symmetric roles of X and R , the theorem is proved. \blacksquare

The results in Theorem 11 are used to obtain bounds on $H(X)$ and $H(R)$ in the following corollary. Define the binary entropy function, $h(\gamma) = -\gamma \log \gamma - (1 - \gamma) \log(1 - \gamma)$ for $0 < \gamma < 1$ and $h(0) = h(1) = 0$.

Corollary 12: Let \mathcal{X} , \mathcal{R} and \mathcal{U} be the respective supports of random variables X , R , and U satisfying (5) – (7) and (63). Then

$$\min\{H(X), H(R)\} \geq h(\pi \lfloor \pi^{-1} \rfloor) + \pi \lfloor \pi^{-1} \rfloor \log \lfloor \pi^{-1} \rfloor \quad (73)$$

$$\geq \log \frac{1}{\pi}, \quad (74)$$

where $\pi = \min_{u \in \mathcal{U}} P_U(u)$ and the right sides of (73) and (74) are equal if and only if π^{-1} is an integer.

Proof: From (67), $\max_{x \in \mathcal{X}} P_X(x) \leq \min_{u \in \mathcal{U}} P_U(u)$. Together with [7, Theorem 10], this establishes (73). To prove (74), we first consider the case when π^{-1} is an integer. Then

$$\begin{aligned} h(\pi \lfloor \pi^{-1} \rfloor) + \pi \lfloor \pi^{-1} \rfloor \log \lfloor \pi^{-1} \rfloor &= h(1) + \pi \pi^{-1} \log \frac{1}{\pi} \\ &= \log \frac{1}{\pi}. \end{aligned}$$

If π^{-1} is not an integer, then

$$1 - \pi \lfloor \pi^{-1} \rfloor < \pi.$$

Hence,

$$\begin{aligned}
& h(\pi \lfloor \pi^{-1} \rfloor) + \pi \lfloor \pi^{-1} \rfloor \log \lfloor \pi^{-1} \rfloor \\
&= \pi \lfloor \pi^{-1} \rfloor \log \frac{1}{\pi \lfloor \pi^{-1} \rfloor} + (1 - \pi \lfloor \pi^{-1} \rfloor) \log \frac{1}{1 - \pi \lfloor \pi^{-1} \rfloor} + \pi \lfloor \pi^{-1} \rfloor \log \lfloor \pi^{-1} \rfloor \\
&> \pi \lfloor \pi^{-1} \rfloor \log \frac{1}{\pi} + (1 - \pi \lfloor \pi^{-1} \rfloor) \log \frac{1}{\pi} \\
&= \log \frac{1}{\pi}.
\end{aligned}$$

Furthermore, the right hand sides of (73) and (74) are equal only if π^{-1} is an integer. This proves the lower bounds on $H(X)$. Due to the symmetric roles of X and R , the theorem is proved. ■

Suppose P_U is not uniform so that $\min_{u \in \mathcal{U}} P_U(u) < |\mathcal{U}|^{-1}$. In this case, (74) shows that

$$\min\{H(X), H(R)\} > \log |\mathcal{U}|. \quad (75)$$

Comparing with (11) and (12) in Theorem 1, a larger initial key requirement and a larger number of channel uses are required for systems which achieve the minimal expected key consumption. The following theorem shows that the lower bounds in (74) can be achieved for certain P_U including the uniform distribution and D -adic distributions, $P_U(u) = D^{-i}$ for certain integers D and i .

Theorem 13: Let $\mathcal{U} = \{1, \dots, \ell\}$ and let $P_U(\ell) \leq P_U(i)$ for $1 \leq i \leq \ell$. If there exists a set of positive integers $\Psi = \{\psi_i\}$ such that $P_U(i) = \psi_i P_U(\ell)$ for $1 \leq i \leq \ell$, then the partition code $\mathcal{C}(\Psi)$ simultaneously achieves the minimum $H(X)$ and $H(R)$ among all EPS systems achieving minimal key consumption.

Proof: Suppose (R, U, X) satisfies (5) – (7) and (63) so that $H(X) \geq \log \frac{1}{P_U(\ell)}$ from (74). Note that $P_U(\ell) = (\sum_{i=1}^{\ell} \psi_i)^{-1}$ from the definition of Ψ . Therefore

$$H(X) \geq \log \left(\sum_{i=1}^{\ell} \psi_i \right). \quad (76)$$

The partition code $\mathcal{C}(\Psi)$ has $\theta = \sum_{i=1}^{\ell} \psi_i$ so that it can achieve equality in (76) from (64). Similarly, we can argue that the partition code $\mathcal{C}(\Psi)$ achieves the minimum $H(R)$. ■

For some other source distributions P_U , the partition code may not achieve the minimal number of channel uses $H(X)$, as illustrated in the following example.

Example 4: Consider an EPS system (R, U, X) such that

- 1) U is a binary random variables where $P_U(0) = 3/5$.

- 2) X and R take values from the set $\{0, 1, 2, 3\}$.
- 3) $P_X(0) = P_R(0) = 2/5$, $P_X(i) = P_R(i) = 1/5$ for $i = 1, 2, 3$.
- 4) $I(X; R) = 0$ so that $P_{XR}(xr) = P_X(x)P_R(r)$ for all x and r .
- 5) U is a function of (X, R) such that $U = 0$ if and only if (i) $X = 0$ and $R \neq 0$, or (ii) $R = 0$ and $X \neq 0$, or (iii) $X = R \neq 0$. Consequently, $P_{U|XR}(u | x, r)$ is well-defined.

It is straightforward to check that $\{U, X, R\}$ satisfies (5) – (7) and (63) and $H(X) = H(R) < \log 5$. However $\theta = 5$ is the smallest integer such that $\theta \cdot P_U(u)$ is an integer. In this example, $H(X)$ is smaller than the value given in (64). While Theorem 13 shows that partition code can simultaneously minimize $H(X)$ and $H(R)$ under the conditions (5) – (7) and (63), this example shows that partition code is not necessarily optimal in terms of minimizing $H(X)$ for a general source.

B. Minimal number of channel uses

In the previous subsection, we proposed partition codes $\mathcal{C}(\Psi)$ which minimize the expected key consumption for error free perfect secrecy systems. However, we also demonstrated that these codes do not guarantee the minimal number of channel uses $H(X)$, among all other EPS systems which also minimize the expected key consumption. Finding an EPS system which minimizes the number of channel uses for a given expected key consumption is a very challenging open problem. In this subsection, we aim to minimize $I(R; UX)$ in the regime where $H(X)$ meets the lower bound in Theorem 1, $H(X) = \log |\mathcal{U}|$. Unlike in Section IV-A, we can completely characterize this regime.

Using Theorem 3, we can show that by using one-time pad,

$$H(U) \leq \log |\mathcal{U}| = H(X) = H(R) = I(R; UX).$$

Therefore, in this instance, the expected key consumption $I(R; UX)$ is not minimal when the source U is not uniform. However, the following theorem shows that among all EPS systems which minimize the number of channel uses, the one-time pad minimizes the expected key consumption.

Theorem 14: Consider any EPS system (R, U, X) (e.g., one-time pad) with $H(X) = \log |\mathcal{U}|$. Then $I(R; UX) = \log |\mathcal{U}|$ and $H(X|RU) = 0$.

Proof: If $H(X) = \log |\mathcal{U}|$, $P_X(x) = 1/|\mathcal{U}|$ for $x \in \mathcal{X}$ and

$$|\mathcal{X}| = |\mathcal{U}| \quad (77)$$

from Theorem 1. Let

$$\mathcal{X}_{ru} = \{x \in \mathcal{X} : P_{RUX}(r, u, x) > 0\}$$

be the set of possible values of X when $R = r$ and $U = u$. Due to (8), $\mathcal{X}_{ri} \cap \mathcal{X}_{rj} = \emptyset$ if $i \neq j$.

Together with (77),

$$|\mathcal{U}| = |\mathcal{X}| \geq \left| \bigcup_u \mathcal{X}_{ru} \right| = \sum_u |\mathcal{X}_{ru}| \geq |\mathcal{U}| \min_u |\mathcal{X}_{ru}|. \quad (78)$$

On the other hand, for any $r \in \mathcal{R}$ and $u \in \mathcal{U}$

$$\sum_{x \in \mathcal{X}_{ru}} P_{RUX}(r, u, x) = P_{UR}(u, r) = P_U(u)P_R(r) > 0$$

from (7), and hence, $|\mathcal{X}_{ru}| \geq 1$. Substituting this result into (78) shows that $|\mathcal{X}_{ru}| = 1$. Therefore, X is a function of R and U , which verifies

$$H(X | UR) = 0. \quad (79)$$

Together with (5) – (7), it is easy to verify that $I(R; UX) = H(X) = \log |\mathcal{U}|$. ■

C. The fundamental tradeoff

An important open problem is to find coding schemes which can achieve points on the tradeoff curve between Points 1 and 2 in Figure 8. For a given source distribution P_U and number of channel uses $H(X) = \log |\mathcal{U}| + \gamma$, with $\gamma \geq 0$ we need to solve the following optimization problem,

$$f(\gamma) = \inf_{P_{RX|U} \in \mathcal{P}_\gamma} I(R; UX), \quad (80)$$

where

$$\mathcal{P}_\gamma = \{P_{RX|U} : I(R; U) = I(X; U) = H(U|X, R) = 0, H(X) = \log |\mathcal{U}| + \gamma\} \quad (81)$$

is the set of feasible conditional distributions yielding an EPS system with the specified number of channel uses.

Solving (80) remains open in general, however two important structural properties of $f(\gamma)$ are given in the following theorem.

Proposition 15: Let P_U and $\gamma \geq 0$ be given. Then \mathcal{P}_γ defined in (81) is non-empty for $\gamma \geq 0$, and $f(\gamma)$ defined in (80) is non-increasing in γ .

Proof: A non-vacuous feasible set is demonstrated as follows. Let (R, U, X) be a given EPS system. Define a second EPS system (R', U', X') as follows. Let $(R', U') = (R, U)$ and $X' = (X, A)$, where A is a random variable independent of (R, U, X) such that $H(A) = \delta$ for any given $\delta \geq 0$. In other words, (R', U', X') is constructed by adding some spurious randomness into the ciphertext of the EPS system (R, U, X) . Setting $\delta = \gamma$ and supposing that (R, U, X) is a cipher system using a one-time pad yields $P_{R'X'|U'} \in \mathcal{P}_\gamma$.

By the same trick, we can show that $f(\gamma)$ is non-increasing. For any $\gamma > 0$ and $\epsilon > 0$, let (R, U, X) be an EPS system such that $P_{RX|U} \in \mathcal{P}_\gamma$ and

$$I(R; UX) < f(\gamma) + \epsilon. \quad (82)$$

It is easy to check that $P_{R'X'|U'} \in \mathcal{P}_{\gamma+\delta}$ and $H(X | UR) = H(X' | U'R') - \delta$. Then

$$f(\gamma + \delta) = \inf_{P_{\tilde{R}\tilde{X}|\tilde{U}} \in \mathcal{P}_{\gamma+\delta}} I(\tilde{X}; \tilde{U}\tilde{R}) \quad (83)$$

$$= \inf_{P_{\tilde{R}\tilde{X}|\tilde{U}} \in \mathcal{P}_{\gamma+\delta}} \left(H(\tilde{X}) - H(\tilde{X} | \tilde{U}\tilde{R}) \right) \quad (84)$$

$$= \log |\mathcal{U}| + \gamma + \delta - \sup_{P_{\tilde{R}\tilde{X}|\tilde{U}} \in \mathcal{P}_{\gamma+\delta}} H(\tilde{X} | \tilde{U}\tilde{R}) \quad (85)$$

$$\leq \log |U| + \gamma + \delta - H(X' | U'R') \quad (86)$$

$$= H(X) - H(X | UR) \quad (87)$$

$$< f(\gamma) + \epsilon, \quad (88)$$

where (88) follows from (82). Since $\epsilon > 0$ is arbitrary, the second claim of the proposition is proved. ■

V. COMPRESSION BEFORE ENCRYPTION

In Section I we discussed the standard approach of compression-before-encryption (cf. Fig. 2) suggested by Shannon. In the following, we will show that this approach is not necessarily the right way to minimize either $I(R; UX)$ or $H(X)$ in error free perfect secrecy systems. For simplicity, all units in this section are in bits and logarithms are with base 2.

A central idea in lossless data compression is to encode frequently occurring symbols (or strings) using shorter codewords. However, this can cause problems in the context of EPS systems. For instance, suppose our cipher consists of a Huffman code followed by a one-time pad using a key with the same length as the Huffman codeword. At first glance, this approach can reduce both the ciphertext size and the key size to the minimum expected codeword length. Unfortunately, this method is not secure because the *length* of the output discloses some information about the message. Consider an extreme case that the message is generated according to $P_U(i) = 2^{-i}$ for $1 \leq i < \ell$ and $P_U(\ell) = 2^{-(\ell-1)}$. If a binary Huffman code is used, the message is uniquely identified by the length when $U < \ell - 1$.

This problem can be solved by different methods. One solution has been discussed in Example 2. In this section, we only consider the *compress-encrypt-pad scheme* of Fig. 9, since this is sufficient to illustrate the deficiencies of compression before encryption.

In Fig. 9, a prefix code is used to encode the message U and a codeword with length $\sigma(U)$ is obtained. The codeword is further encrypted by one-time pad using a key with the same length $\sigma(U)$. After application of the one-time pad, fair bits are appended such that the output has a constant length γ equal to the longest codeword, $\max_{u \in \mathcal{U}} \sigma(u)$. The receiver decrypts the message by applying the key bit-by-bit to the ciphertext until a codeword in the prefix code is obtained.

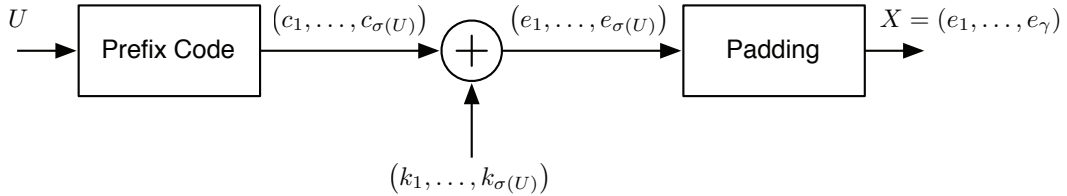


Fig. 9. A compression-encryption-padding scheme.

In this scheme, the ciphertext X has a uniform distribution so that $H(X) = \gamma$. Since γ is the length of the longest codeword and a prefix code is uniquely decodable, $\gamma \geq \log \ell$, where $\ell = |\mathcal{U}|$. Therefore, $H(X) \geq \log \ell$, in agreement with Theorem 1. This scheme requires an initial key of length $H(R) \geq \log \ell$ bits providing a sufficiently long secret key in case the longest codeword is the one that happens to be generated.

Let us now compare the performance of this scheme with the bounds obtained in Section IV-A, where the minimal expected key consumption is assumed. Suppose the Shannon code [4] is used in the scheme described in Fig. 9 to construct an EPS system. The performance is given in the following theorem.

Theorem 16: If the Shannon code is used in the compress-encrypt-pad scheme described in Fig. 9 to construct an EPS system, then

$$H(R) = H(X) = \left\lceil \log \frac{1}{\pi} \right\rceil, \quad (89)$$

which exceeds the lower bound in (74) by no more than 1 bit. Furthermore, the expected key consumption exceeds the lower bound (56) by no more than 1 bit,

$$I(R; UX) \leq H(U) + 1.$$

Proof: Recall that $\sigma(u)$ is the length of the codeword assigned to $U = u$. Then the longest codeword has length equal to

$$\left\lceil \log \frac{1}{\pi} \right\rceil, \quad (90)$$

where $\pi = \min_{u \in \mathcal{U}} P_U(u)$. Recall in Fig. 9 that fair bits are appended to each codeword to construct a constant length ciphertext X . Therefore, $H(R) = H(X) = \left\lceil \log \frac{1}{\pi} \right\rceil$ which is within one bit of the lower bound in (74). Furthermore, the expected key consumption

$$I(R; UX) = H(R) - H(R | UX) \quad (91)$$

$$= \left\lceil \log \frac{1}{\pi} \right\rceil - \left(\left\lceil \log \frac{1}{\pi} \right\rceil - \sum_{u \in \mathcal{U}} P_U(u) \sigma(u) \right) \quad (92)$$

$$= \sum_{u \in \mathcal{U}} P_U(u) \sigma(u) \quad (93)$$

$$\leq H(U) + 1, \quad (94)$$

where (92) follows from the fact that $H(R | UX)$ is equal to the number of appended fair bits, and (94) follows from [4, (5.29)–(5.32)]. Therefore, $I(R; UX)$ is also within a bit of the lower bound in (56). ■

Therefore, we conclude that if the Shannon code is used for compression in Fig. 9, then the performance is close to the optimal code in the minimal key consumption regime when both $H(U) \gg 1$ and $\log \frac{1}{\pi} \gg 1$.

Now, we compare the performance obtained when the Huffman code is used in place of the Shannon code. In this case, the expected key consumption $I(R; UX)$ can again be analyzed similar to (91) – (94). Since the expected codeword length in (93) is shorter for the Huffman code, a smaller $I(R; UX)$ can be obtained. However, the longest codeword in the Huffman code can be longer than the longest codeword in the Shannon code. As a consequence, larger $H(X)$ and $H(R)$ are required for certain P_U . This can be seen in the example in Table I. In the worst case, the longest codeword in the Huffman code can be as much as 44% longer than the longest codeword in the Shannon code [10]. Furthermore, the partition code $\mathcal{C}(\Psi)$ in Table I outperforms the compression before encryption schemes based on either the Huffman code or the Shannon code because $\mathcal{C}(\Psi)$ is optimal according to Theorem 13. On the other hand, the Shannon code uses unnecessarily long codewords for certain source distributions, e.g., $P_U = (0.9, 0.1)$. As a consequence, larger $H(X)$ is needed as shown in Table II. However, the minimal $I(R; UX)$ or the minimal $H(X)$ can be obtained using different partition codes. We conclude that compression before encryption is a suboptimal strategy to minimize key consumption or the number of channel uses in EPS systems.

TABLE I
COMPARING DIFFERENT SCHEMES WITH $\Phi = (1, 1, 1, 3, 4, 7, 11)$ AND $P_U(i) = \Phi(i)/28$ FOR $1 \leq i \leq 7$

	Huffman	Shannon	Partition $\mathcal{C}(\Phi)$
$I(R; UX)$	2.357	2.679	$2.291 = H(U)$
$H(X)$	6	5	5

TABLE II
COMPARING DIFFERENT SCHEMES WITH $\Phi = (9, 1)$, $\Phi' = (1, 1)$ AND $P_U = (0.9, 0.1)$

	Huffman	Shannon	Partition $\mathcal{C}(\Phi)$	Partition $\mathcal{C}(\Phi')$
$I(R; UX)$	1	1.3	$0.469 = H(U)$	1
$H(X)$	1	4	4	1

Suppose now that the source distribution is d -adic and the smallest probability mass in P_U is equal to d^{-k} for certain integers d and k . What were binary digits in the scheme described above in Fig. 9 now become d -ary symbols. It can be verified that the longest codeword has length equal to k . Therefore, both d -ary Shannon codes and d -ary Huffman codes can achieve the minimal

$H(X)$ and $H(R)$ in (74). Furthermore, the expected codeword length is equal to $H(U)$. By (93), $I(R; UX)$ is equal to the expected codeword length, which is equal to $H(U)$. Therefore, the minimal $I(R; UX)$ is achieved. However, a prefix code cannot achieve the expected codeword length $H(U)$ when P_U is not d -adic [5, Theorem 4.6]. Again consider the example in Table II where $P_U = (0.9, 0.1)$. Only partition code but neither the Shannon nor the Huffman code can be used to achieve $I(R; UX) = H(U)$. Indeed, the d -adic distribution is just a special case of the condition used in Theorem 13. Therefore, the partition code can achieve the minimal $I(R; UX)$ for a wider range of P_U .

VI. PROOF OF THEOREM 9

Suppose there exists $u \in \mathcal{U}$ such that $P_U(u)$ is irrational. Define a new random variable U^* such that

$$U^* = \begin{cases} 0 & \text{if } U = u \\ 1 & \text{otherwise.} \end{cases}$$

Then $P_{U^*}(0)$ and $P_{U^*}(1)$ are irrational. As U^* is a function of U , by (5) – (7) and (63),

$$I(U^*; R) = I(U^*; X) = I(X; R) = H(U^* | XR) = 0. \quad (95)$$

Therefore, it suffices to consider binary U .

Let \mathcal{X} and \mathcal{R} be the respective supports of X and R . Suppose to the contrary first that $|\mathcal{X}|$ and $|\mathcal{R}|$ are both finite. We can assume without loss of generality that

$$\mathcal{X} = \{1, \dots, n\} \quad (96)$$

$$\mathcal{R} = \{1, \dots, m\}. \quad (97)$$

Let

$$x_i = P_X(i), \quad i = 1, \dots, n \quad (98)$$

$$r_j = P_R(j), \quad j = 1, \dots, m, \quad (99)$$

and let \mathbf{x} be the n -row vector with entries x_i . Similarly, define the column vector \mathbf{r} .

As X and R are independent and $H(U | XR) = 0$, there exists a function g such that $U = g(X, R)$. Hence, from X and R we induce a $n \times m$ *decoding matrix* G with entries

$$G_{i,j} = f(i, j), \quad i = 1, \dots, n, j = 1, \dots, m.$$

Then

$$\sum_{j=1}^m G_{i,j} r_j = P_U(1), \quad i = 1, \dots, n \quad (100)$$

$$\sum_{i=1}^n x_i = \sum_{j=1}^m r_j = 1 \quad (101)$$

$$x_i \geq 0, r_j \geq 0, \quad i = 1, \dots, n, j = 1, \dots, m \quad (102)$$

$$\sum_{i=1}^n x_i G_{i,j} = P_U(1), \quad j = 1, \dots, m \quad (103)$$

Here, (100) is due to the fact that $I(U; X) = 0$, (101) and (102) are required since P_X and P_R are probability distributions, and (103) follows from $I(U; R) = 0$.

In fact, for any \mathbf{x} , \mathbf{r} and binary matrix G satisfying the above four conditions, one can construct random variables $\{U, R, X\}$ such that

$$I(U; R) = I(U; X) = I(X; R) = H(U | XR) = 0 \quad (104)$$

where $U = f(X, R)$ and the probability distributions of X and R are specified by the vectors \mathbf{x} and \mathbf{r} respectively.

In the following, we will prove that if the rows of G are not independent, then we can construct another random variable X^* with support \mathcal{X}^* , $|\mathcal{X}^*| < |\mathcal{X}|$ such that

$$I(U; R) = I(U; X^*) = I(X^*; R) = H(U | X^* R) = 0. \quad (105)$$

To prove this claim, suppose that there exists disjoint subsets \mathcal{A} and \mathcal{B} of $\{1, \dots, n\}$ and positive numbers $\alpha_i, i \in \mathcal{A} \cup \mathcal{B}$ such that

$$\sum_{i \in \mathcal{A}} \alpha_i G_i = \sum_{k \in \mathcal{B}} \alpha_k G_k. \quad (106)$$

where G_i is row i of G . Then we will claim that

$$\sum_{i \in \mathcal{A}} \alpha_i = \sum_{k \in \mathcal{B}} \alpha_k.$$

Multiplying both sides of (106) by \mathbf{r} ,

$$\sum_{i \in \mathcal{A}} \alpha_i G_i \mathbf{r} = \sum_{k \in \mathcal{B}} \alpha_k G_k \mathbf{r} \quad (107)$$

$$\sum_{i \in \mathcal{A}} \alpha_i P_U(1) = \sum_{k \in \mathcal{B}} \alpha_k P_U(1) \quad (108)$$

$$\sum_{i \in \mathcal{A}} \alpha_i = \sum_{k \in \mathcal{B}} \alpha_k. \quad (109)$$

Let $\epsilon \triangleq \min_{i \in \mathcal{A} \cup \mathcal{B}} x_i / \alpha_i$. Assume without loss of generality that $n \in \mathcal{A}$ and that $\epsilon = x_n / \alpha_n$. Define

$$x_i^* = \begin{cases} x_i - \epsilon \alpha_i, & i \in \mathcal{A} \\ x_i + \epsilon \alpha_i, & i \in \mathcal{B} \\ x_i, & \text{otherwise.} \end{cases}$$

Note that $x_n^* = 0$. Suppose that the probability distribution of X is changed such that $P_X(i) = x_i^*$. Then it can be checked easily that U, X, R still satisfy (5) – (7) and (63). Furthermore, the size of the support of P_X is $|\mathcal{X}| \leq n - 1$.

Repeating this procedure, we can prove that for any random variable U , if there exists auxiliary random variables X, R satisfying (5) – (7) and (63), then there exists auxiliary random variables X^*, R^* such that (95) is satisfied and the rows and columns of the decoding matrix induced by X^* and R^* are all linearly independent. Hence, the decoding matrix G induced by X^* and R^* must be square (and thus $m = n$). Consequently,

$$\sum_{i=1}^n x_i G_{i,j} = P_U(1), \quad j = 1, \dots, n. \quad (110)$$

There exists a unique solution (z_1, \dots, z_n) such that

$$\sum_{i=1}^n z_i G_{i,j} = 1, \quad j = 1, \dots, n. \quad (111)$$

Clearly, $z_i = x_i / P_U(1)$. As all the entries in G are either 0 or 1, all the z_i are rational numbers. Therefore,

$$1 = \sum_{i=1}^n x_i = P_U(1) \sum_{i=1}^n z_i. \quad (112)$$

Hence, $P_U(1)$ must be rational and a contradiction occurs. We have proved that \mathcal{X} and \mathcal{R} cannot be both finite. The case when only \mathcal{X} or \mathcal{R} is finite can be similarly proved.

VII. CONCLUSION

This paper studied perfect secrecy systems with zero decoding error at the receiver, with the additional assumption that the message U and the secret key R are independent, $I(U; R) = 0$. Under this setup, we found a new bound $\log |\mathcal{U}| \leq H(R)$ on the key requirement, improving on Shannon's fundamental bound $H(U) \leq H(R)$ for perfect secrecy.

To transmit the ciphertext X , the lower bound on the minimum number of channel uses has been shown to be $\log |\mathcal{U}| \leq H(X)$. If the source distribution is defined on a countably infinite support or a support with unbounded size, no security system can simultaneously achieve perfect secrecy and zero decoding error.

We also defined and justified three new concepts: *residual secret randomness*, *expected key consumption*, and *excess key consumption*. We have demonstrated the feasibility of extracting residual secret randomness in multi-round secure communications which use a sequence of error free perfect secrecy systems. We quantified the residual secret randomness as $H(R|UX)$. We further distinguished between the size $H(R)$ of the secret key required prior to the commencement of transmission, and the expected key consumption $I(R; UX)$ in a multi-round setting. In contrast to $H(R) \geq \log |\mathcal{U}|$, we showed that $I(R; UX)$ is lower bounded by $H(U)$, giving a more precise understanding about the role of source entropy in error free perfect secrecy systems. The excess key consumption is quantified as $I(R; X)$, and is equal to 0 if and only if the minimal expected key consumption is achieved.

One of the main objectives of this paper was to reveal the fundamental tradeoff between expected key consumption and the number of channel uses. For the regime where the minimal $I(R; UX)$ is assumed, $H(X)$ and $H(R)$ are inevitably larger and corresponding lower bounds for $H(X)$ and $H(R)$ have been obtained. If the source distribution P_U has irrational numbers, the additional requirements on the alphabet sizes of X and R to achieve minimal $I(R; UX)$ have been shown. We have proposed a new code, the *partition code*, which generalizes the one-time pad, and can achieve minimal $I(R; UX)$ when all the probability masses in P_U are rational. In some cases, the partition code can simultaneously attain the minimal $H(X)$ and $H(R)$ in this regime.

At the other extreme, the regime where the minimal number of channel uses is assumed, the one-time pad has been shown to be optimal. For the intermediate regime, we have formulated an optimization problem for the fundamental tradeoff between $I(R; UX)$ and $H(X)$. We also demonstrated that compression before encryption cannot minimize either $H(R)$, $H(X)$ or $I(R; UX)$.

This paper has highlighted a few open problems. First, the complete characterization of the tradeoff between $I(R; UX)$ and $H(X)$ remains open. Second, the partition code is only one class of codes designed to minimize expected key consumption. Codes achieving other points

on the tradeoff curve are yet to be discovered. In particular, a code achieving minimal $H(X)$ and $H(R)$ in the regime of minimal expected key consumption is important for the design of efficient and secure systems.

REFERENCES

- [1] C. E. Shannon, "Communication Theory of Secrecy Systems," *Bell Syst. Tech. J.*, vol. 28, pp. 656–715, Oct. 1949.
- [2] J. L. Massey, "An introduction to contemporary cryptology," *Proc. IEEE*, vol. 76, pp. 533–549, May 1988.
- [3] S.-W. Ho, "On the interplay between Shannon's information measures and reliability criteria," *IEEE Int. Symp. Inform. Theory*, (Seoul, Korea), 28 June-3 July, 2009.
- [4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 2 Ed., 2006.
- [5] R. W. Yeung, *Information Theory and Network Coding*, Springer, 2008.
- [6] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, 1979.
- [7] S.-W. Ho and S. Verdú, "On the interplay between conditional entropy and error probability," *IEEE Trans. Inform. Theory*, vol. 56, pp. 5930–5942, Dec 2010.
- [8] G. Vernam, "Cipher printing telegraph systems for secret wire and radio telegraphic communications," *Trans. American Inst. Elec. Eng.*, vol. XLV, pp. 295–301, 1926.
- [9] T. H. Chan and S.-W. Ho, "2-Dimensional Interval Algorithm," *IEEE Inform. Theory Workshop*, (Paraty, Brazil), pp. 633–637, Oct. 2011.
- [10] Y. Abu-Mostafa and R. McEliece, "Maximal codeword lengths in Huffman codes," *Computers & Mathematics with Applications*, vol. 39, no. 11, pp. 129–134, 2000.